# Specific Aims

Schizophrenia (SCZ) is a severe mental disorder that affects millions of people worldwide. In recent years, multiple genome-wide association studies (GWASs) have made substantial progress, identifying several promising candidate genes with common variants.1-6 However, like many other complex diseases, all these candidates only account for a small proportion of observed heritability.7-10 Many more candidates remained to be discovered due to the small effects of individual variants/genes. More recently, sequencing and copy number variation (CNV) analyses have documented many rare de novo mutations and insertion/deletions (InDels) that predispose to risks to this disorder.11-24 These studies provide strong evidence that CNV and rare variants, including de novo variants, contribute significantly to genetic risks to SCZ, and collectively, they are an important component of the genetic architecture of SCZ. Combined with what we have learned from GWASs, it is clear that both common and rare variants contribute to genetic risks to SCZ, and they form a continuum of genetic risks.25-27

Our current knowledge of SCZ is overwhelmingly derived from the study of Caucasian populations. The understudy of other ethnicities and the lack of systematic examination of the differences between populations are noticeable in the field. These weaknesses may impede our understanding of the etiology of the disorder. In responding to the request for application AI-12-021 “U.S.-China Program for Biomedical Collaborative Research (R01)”, we propose studies aiming at the understanding of genetic architecture of SCZ in the Chinese population and investigating the shared and ethnic-specific risk factors between the Chinese and Caucasian populations. Our aims are the following:

1. Conduct exome sequencing for 140 nuclear Chinese families with multiple affected individuals to discover single nucleotide variations (SNVs) and copy number variations (CNVs) predisposing to the disorder. We plan to use families with both parents (affected or unaffected), 2 affected and 1 unaffected siblings and families with 1 parent, 2 affected and 1 unaffected siblings. The use of families with affected and unaffected siblings allows us to simultaneously discover and characterize transmitted and de novo risk variants. The use of unaffected siblings in the families serves as a better control than subjects from general population, as it can filter out many non-causal variants observed in the families.
2. Genotype 5,000 cases and 5,000 controls for up to 100 the most promising risk variants discovered in Aim 1 above to verify their association with SCZ. We will apply a set of sophisticated statistical, bioinformatic and functional filters to select the most promising SNVs. We will focus on those rare variants (including de novo variants) with potential functional consequences and variants occurred at multiple sites in the same genes and biological pathways.
3. Perform comparative analyses using GWAS datasets from both Caucasian and Chinese population to estimate the overlap of genetic risk between the two populations, and to discover and characterize the shared- and ethnic-specific risk genes. We propose to use polygenic analyses to examine the correlation between the published PGC and Chinese GWASs to estimate the overlap of risk factors between these populations, and to examine the genetic structure of SCZ in these two ethnic groups.

The goal of this application is to understand the genetic structure of SCZ in the Chinese population, with a focus on rare and de novo variants segregated in the families, and shared and ethnic specific genetic factors between the Chinese and Caucasian populations. The application takes advantage of next-generation sequencing (NGS) techniques and multiplex Chinese family samples to simultaneously discover both transmitted and de novo risk variants and verifies the promising variants with large independent samples. This is an economically efficient and scientifically novel design. Based on the examination of study designs aiming at identification of rare risk variants,28,29 where the use of family-history positive cases leads to increase of statistic power in discovery of rare risk variants, our design will led to discover and identify new risk factors and shed light on our understanding of the genetic architecture of SCZ.

# Significance

Genetic studies of SCZ have made significant progress in recent years and multiple loci have been identified in GWASs. Like many other complex diseases, these loci with common variants account for only a small proportion of the total heritability,7-9 and few causal variants in these loci have been found. There may be several factors to explain the missing heritability. GWAS is based on the premise that common diseases are caused by common variants. If the premise is true and we still cannot account for the majority of heritability, it would suggest that there may be substantially more risk genes with small effects. The reason that these genes have not been identified is likely due to lack of power in the GWASs.27,30 But if many cases of SCZ is not caused by common variants, we would not expect GWAS to be very successful for identifying the most of genetic risks. Whole genome sequencing and exome sequencing, on the other hand, is based on a different premise that single common and rare variant, or combinations of them, are the major cause of the disorder. Recent studies have provided strong evidence that rare variants, including CNVs and de novo mutations, play an important role in SCZ.11-24 This is consistent with a recent survey of GWASs in which both common and rare variants contribute to disease risks, although the extent to which common and rare variants contribute varies across diseases.25 Since the genetic architecture of SCZ is likely to involve both common and rare variants, to ensure successful identification of risk factors, a better study design needs to consider both common and rare variants and accommodates their varying contributions. In this application, we propose to conduct exome sequencing of multiplex families with parents, affected and unaffected siblings. This design enables us to detect both common and rare risk variants, including de novo rare variants. In addition, our team (Dr. Xiong of UT Health Center at Houston) has recently developed novel strategies31,32 to jointly test both common and rare variants unifying family and population study desing (Zhu and Xiong, 2012). With our family design and the application of these new analytic strategies, we could significantly increase our power to detect the real effects of genes. The unified approach can also correct for unknown population stratification, family structure and cryptic relatedness while maintaining high power in the sequence-based association studies

SCZ, like many other complex diseases, can have sporadic cases and cases run in families. Early studies of cases run in families, also referred to as cases with family history (FH), have higher genetic loading, and potentially more homogeneous genetically. [ref] However, we know very little of the nature of FH in SCZ. What constitutes a FH in SCZ? Is there a difference in genetic architecture between cases with and without FH? Most recent studies, including GWASs, do not make a distinction between cases with and without FH. A recent study tried to model family history in a GWAS but failed.33 A rare CNV was found to be associated with male patients with FH.13 A population linkage study supported a model of rare but highly penetrant disease alleles in cases with FH.34 These seem to imply that rare variants are more likely than common variants to explain FH. In this study, we plan to sequence cases with FH exclusively. Because our design is intended to find rare variants, our study could provide novel insights for understanding the nature of these cases with FH, substantially complementing and improving our knowledge of the genetic architecture of SCZ.

SCZ occurs in all ethnicities with similar prevalence,35 and most of our understanding of SCZ comes from the studies of Caucasian populations. There is evidence that some of the factors influencing the development of SCZ are common across different ethnic groups and others are ethnic specific. [ref] However, we know very little of these shared and ethnic specific factors. Since our Chinese team has finished a large GWAS recently, we have a unique opportunity to examine and compare the genetic factors between the Chinese and Caucasian populations, which will significantly increase our power to unravel the genetic structure of SCZ.

# Innovation

Recent GWASs, CNV analyses and sequencing have provided ample evidence that both common and rare variants influence the development of SCZ. The common variants were identified by GWASs. The rare variants, including copy number variations and de novo mutations, were found in GWASs and sequencing of trio families. Our current knowledge of the genetics of SCZ is overwhelmingly derived from the studies of Caucasian population. Lack of study of the genetic architecture in other ethnicities is an obvious weakness in our understanding of the disease. Are there shared- and ethnic-specific genetic liabilities across different ethnic groups? If yes, what are these share- and ethnic specific factors, and how they contribute to the architecture of the disorder? Another area with relatively few studies is the family history of SCZ. Is there a difference between sporadic cases and those cases from multiplex families in genetic architecture? This application proposes studies to examine these questions, its innovations include:

1. We take a unique multiplex family design that includes parents, affected and unaffected offspring for exome sequencing. The exclusive inclusion of pedigrees with family history of SCZ which gives us the opportunity to examine family history related risk factors, and comparing these factors with that discovered from sporadic cases in the literature is highly innovative. This will lead to a better understanding of the genetics of SCZ. Cases with positive family history are also likely more homogeneous, using only these cases would increase our power to detect risk variants enriched in these multiplex families.
2. Inclusion of parents and an unaffected sibling in the pedigrees which allows us to discover and identify risk de novo mutations more reliably is a novel approach to genetic studies of SCZ. Unaffected siblings share about 50% genetic materials with affected subjects, they serve as better controls than the general population for rare and de novo risk variant analyses.
3. Application of newly developed statistics to jointly test common and rare variants discovered in our sequencing. It has been known that rare variants have lower power for association testing due to their low frequencies. One of our investigators, Dr. Momiao Xiong, has recently developed a strategy to simultaneously analyze multiple variants across a genomic interval in both family and population data. The ability of the smoothed functional principal component analysis (SFPCA) to test association of both common and rare variants offers an innovative approach to boost power and improve flexibility of our association analysis.
4. There is evidence that there are shared- and ethnic-specific genetic factors in SCZ. With the availability of large GWA datasets from both the Chinese and Caucasian populations, we creatively use this unique opportunity to study and identify these shared- and ethnic-specific risk factors. To our knowledge, this will be the first study to systematically examine this issue.

# Approach

## The US-China collaboration team, responsibility and publication policy

In response to this RFA, we have selected researchers from both US and Chinese institutions to form a collaborative team to work together for the aims listed above. The US side includes Drs. Xiangning Chen and Kenneth Kendler of Virginia Commonwealth University, and Dr. Momiao Xiong at the University of Texas Health Science Center at Houston. Drs. Chen and Kendler both serve as principle investigators for this application. The Chinese side includes Dr. Lin He at the Shanghai Jiao Tong University and … [Lin, please supply other investigators if any]. Dr. He will serve as the principle investigator on the Chinese side.

During the preparation of this application, investigators from both sides have had extensive discussions on the design of studies proposed, and agreed to share the resources necessarily to accomplish the goals. Our plan to share the resources and workload includes:

1. For Aim 1, the US side will be responsible to organize and select the DNA samples used in exome sequencing. Dr. Chen has applied and obtained approval from NIMH official to use the Chinese pedigrees in the NIMH Genetics Initiative Repository (<https://www.nimhgenetics.org/nimh_human_genetics_initiative/>). Of the 700 individuals to be sequenced, each side will use its budget to sequence 350 individuals. While the funds for sequencing are from respective countries, the experimental protocols, data quality controls and analyses will be the same to ensure scientific integrity.
2. For Aim 2, the Chinese team will be responsible to organize and select DNA samples used for verification. Our Chinese team has recently published a GWA study with a combined sample size of 8,123 cases and 11,007 controls.36 We will select 5,000 cases and 5,000 controls from this dataset [Lin, please confirm if we can use all the cases and controls, it would be better if we do since we need power to detect these rare variants] and use it to verify the top candidates selected from genetic analyses of Chinese families with exom sequencing data (Aim 1). Both sides will pool equal amount of money to conduct genotyping of the selected SNVs. Due to the Chinese government regulations on DNA exporting, the replication study will be conducted in China at the site of Dr. He’s lab.
3. For Aim 3, the US team will use the PGC SCZ dataset, and the Chinese team will provide the GWA dataset published recently. Both sides will discuss and design the procedures for the analyses, and split the workload. Since the US team has better computational resources, it will execute the procedures and share the results with Chinese team.

All data produced from this application will be shared between the two teams. All papers generated from this collaboration will be discussed and agreed on before submission to journals for publication. Authorship will be based on individual’s contribution, and be negotiated and agreed on by all authors. Intellectual property produced from this study will be shared and negotiated by the respective institutions.

The US and Chinese teams have built a mutually trusting relationship, and we have worked together on genetics of schizophrenia, and successfully published the results.37,38 Communications between us will be carried out by email, teleconference, google talk. Each year we will have a meeting to plan the design, working schedule, discuss the problems and present scientific findings. We are confident that we will have a productive collaboration.

## Preliminary studies

### Genetic studies of schizophrenia

Our team has extensive experience and expertise in the genetics of schizophrenia. Dr. Kendler is an internationally recognized leader in genetic studies of schizophrenia. He is the recipient of the 2011 International Society of Psychiatric Genetics’ Ming Tsuang Lifetime Achievement Award and 2011 World Psychiatric Association’s Jean Delay Prize. Dr. Kendler’s accomplishments include the establishment of familiar risk in schizophrenia with family and adoption studies in the 1980s,39-43 and contribution to the most recent PGC SCZ GWA studies.4,6,44 Dr. He, member of science acaemy in China, is also a well-established scholar in genetic study of SCZ, and has published many important works in the field, including the discovery of SLIT3,45 CHI3L1,46 BCL947 and synapsin II48 association with SCZ, and testing and confirming the association of many leading candidates49-54 in Chinese populations (see biosketch for more relevant papers). More recently, Dr. He has published a large GWAS,36 identifying 2 risk loci in Chinese population. Dr. Chen has more than 10 years of experience in SCZ studies, and has identified the association of the SPEC2/PDZ-GEF2/ ACSL6 locus,55 MEGF10,56 IL3,57 CSF2RB58 and CMYA559 genes with SCZ.

### Next generation DNA sequencing and analysis

**RNA sequencing:** Our team has direct working experience with data produced from the next generation DNA sequencing technologies. In one study, we conducted transcriptome sequencing for RNA samples isolated from blood samples of SCZ patients and healthy controls. In this study, we first sequenced 3 cases and 3 controls, followed with verification of sequencing two pools of 10 cases and 10 controls. We found that there were 198 genes differentially expressed between cases and controls, of them 19 had been verified by the pooled sequencing dataset and 21 reached nominal significance in gene-based association analyses of the Molecular Genetics of SCZ GWA dataset. Two genes, S100A8 and TYROBP, had consistent changes in expression in both individual and pooled sequencing datasets and were nominally significant in gene-based association analysis.60

Dr. Chen, one of the PIs in this application, has an ongoing transcriptome sequencing project funded by NARSAD. In that project, we proposed to sequence the transcriptome of post-mortem anterior cingulate cortex (Brodmann Area 24) from patients with SCZ, bipolar disorder (BIP) and controls from the Stanley Medical Research Institute. This is a project in collaboration with Beijing Genomics Institute (BGI). We have finished sequencing and quality assessments for 82 subjects (26 controls, 25 BIP patients and 31 SCZ patients), and aligned the unique reads to human genome reference. On average, for each subject we have 4.7, 12.9 and 10 million unique reads mapped to the human genome for the BIP, SCZ and control groups respectively (Table 1). In RNA-seq data, gene expression is measured by the reads per kilobase exon per million reads (RPKM). We also generated allele specific expression (ASE) data which allow levels of ASE to vary across SNPs in the genes. SNPs from RNA-seq can be identified by consensus alignment of sequencing reads from the pool of multiple individuals.61,62 In this study, we focused on the comparison between SCZ and BIP. We intended to test if there is a correlation in differentially expressed genes between these two disorders since from the GWA studies there is evidence that these two disorders share some common liabilities. We first conducted gene expression analyses using t test, and found evidence that the test statistics (t score) between SCZ and BIP are correlated (Figure 1). We also intended to test if there is an excess of rare variants in those differentially expressed genes. Based on our data, we define rare variants as those observed only once in our sequencing of 82 subjects. After controlling for sequencing depth and total reads mapped to genes, we found that genes differentially expressed (FDR q value cut-off at 0.05 and 0.1) in SCZ patients were enriched in rare variants (Table 2), and the results for BIP were not significant because of lower sequencing coverage. These results are consistent with the observation that multiple rare mutations may be an important source for genetic risk for SCZ.

**Table 1. Summary of RNA sequencing of post-mortem samples**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **N** | **Total Reads** | **Reads Mapped to Genome** | **Reads Mapped to Gene** | **Reads /Gene** | **RPKM /Gene** | **Coverage /Gene** | **SNPs /Subject** | **Genes /Subject** |  |
| Ctrl | 26 | 11,426,942 | 9,999,730 | 5,320,966 | 315 | 24.69 | 0.55 | 58,369 | 15,879 |  |
| BIP | 25 | 5,272,478 | 4,665,049 | 2,314,132 | 144 | 25.92 | 0.47 | 27,820 | 15,294 |  |
| SCZ | 31 | 14,733,231 | 12,894,782 | 6,418,990 | 378 | 23.59 | 0.62 | 68,662 | 16,640 |  |

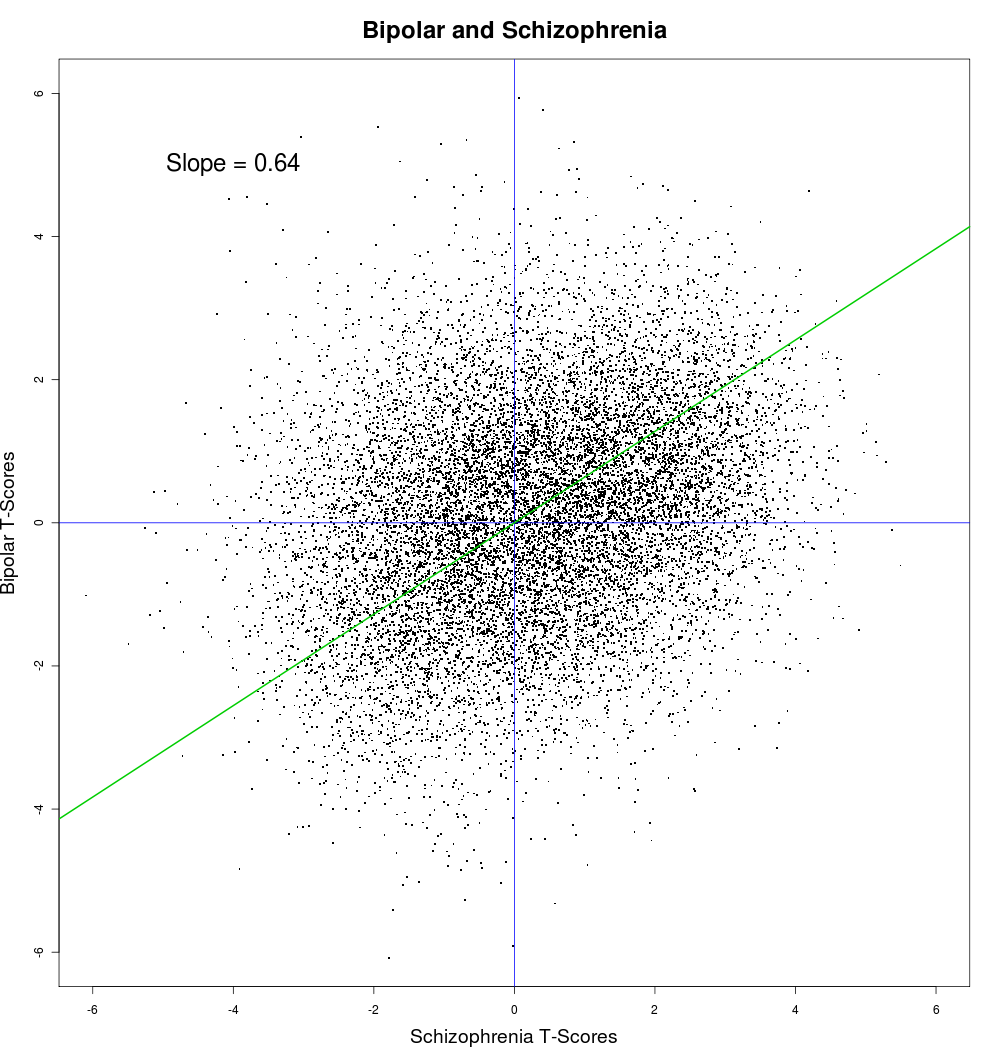


Figure 1. Correlation of gene expression between SCZ and BIP.

**Table 2. Differentially expressed genes in SCZ are enriched for rare variants**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Schizophrenia | | Bipolar disorder | |
| Q-value cut-off (# genes) | Coefficient | P-value | Coefficient | P-value |
| 0.1 (885) | 0.36 | **6x10-7** | -0.24 | 0.85 |
| 0.05 (213) | 0.40 | **0.007** | 0.15 | 0.73 |
| 0.01 (53) | 0.39 | 0.20 | -0.06 | 0.72 |

Can we replace or add the results for ASE analysis? ( The results are attached in my email).

**Pathway association analysis of sequencing data:** We have recently developed the SFPCA statistics to test association of both common and rare variants in a gene and demonstrated that SFPCA is a powerful method to identify association at the level of genes.32 We now extend SFPCA from a single gene to multiple genes and compare the difference in functional principal component scores that are calculated from all genes in a pathway between cases and controls. The SFPCA-based statistic for testing the association of pathway with the disease combines a measure of goodness-of-fit with a roughness penalty to retain the advantages of basis expansion and reduce the dimensionality of the data in the pathway. The SFPCA can utilize merits of both individual variant analysis and group tests. It can also efficiently use information of both risk and protective variants and allow for sign and size heterogeneity of genetic variants in the pathways. There is increasing consensus that complex diseases are caused by common and rare variants. Many statistics can be used to test for association of either common variants or rare variants, but very few can be used to test association of both common and rare variants. The SFPCA is designed to test the association of the entire allelic spectrum of genetic variation.

****

To evaluate the performance of the SFPCA-based statistic for pathway analysis, we will use large scale simulations to calculate the type I error rates and systematically evaluate the power of 23 statistical methods: SFPCA, FPCA,63 the weighted sum (WSS ),64 variable-threshold (VT),65 combined multivariate and collapsing (CMC),66 linear combination test (LCT/LCT),67 quadratic test (QT/QT),67 decoration test (DT/DT),67 WSS/Sidak, WSS/Fisher combination, WSS/Fisher Exact, WSS/GESA,VT/Sidak, VT/Fisher combination, VT/Fisher Exact, VT/GESA, CMC/Sidak, CMC/Fisher combination, CMC/Fisher Exact, CMC/GESA, PCA ( Alan Hulian Lzenman 2008, Modern multivariate statistical techniques: regression, classification, and manifold learning, Springer), SKAT,68 and GESA.69 Using simulation, we demonstrate that SFPCA has appropriate type I error rate (data not shown) and has higher power than other test statistics compared (Figure 2). We applied the SFPCA and other popular statistics for pathway analysis to the early-onset myocardial infarction (EOMI) exome sequence datasets with subjects of European origin (EA, n = 544, 188 cases and 356 controls) and African origin (AA, n = 312, 39 cases and 273 controls) from the NHLBI’s Exome Sequencing Project, and found that VEGF and TGFβ signaling pathways were significantly associated with EOMI in EA population by SFPCA but not by other methods (Table 3). Both VEGF70,71 and TGFβ72,73 signaling pathways have been implicated in cardiovascular diseases and myocardial infarction. These results demonstrate that SFPCA is more powerful than other popular statistics in discovery of association at the level of pathways as well.

**Table 3. P-values associated with EOMI**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Name of Pathway | VEGF signaling pathway | TGF beta signaling pathway |
|  | Number of Genes | 68 | 17 |
|  | SFPCA | 7.26E-06 | 2.00E-05 |
|  | FPCA | 4.17E-03 | 8.30E-04 |
|  | LCT/LCT | 3.41E-01 | 6.93E-03 |
|  | QT/QT | 6.96E-01 | 5.51E-02 |
| EA | DT/DT | 2.45E-01 | 3.92E-01 |
|  | WSS/Fisher Exact | 1.18E-02 | 2.87E-01 |
|  | VT/Fisher Exact | 2.37E-01 | 4.75E-02 |
|  | CMC/Fisher Exact | 4.81E-01 | 5.79E-02 |
|  | CMC | 7.40E-01 | 8.94E-01 |
|  | SKAT | 4.07E-02 | 2.44E-03 |
|  | GSEA | 7.40E-01 | 7.04E-02 |
|  | PCA | 4.40E-03 | 1.20E-03 |
|  | SFPCA | 2.85E-01 | 9.04E-05 |
|  | FPCA | 3.32E-01 | 5.48E-04 |
|  | LCT/LCT | 4.09E-01 | 4.28E-02 |
|  | QT/QT | 1.11E-01 | 1.09E-01 |
| AA | DT/DT | 6.20E-01 | 1.91E-01 |
|  | WSS/Fisher Exact | 2.04E-01 | 2.61E-01 |
|  | VT/Fisher Exact | 3.98E-01 | 2.62E-01 |
|  | CMC/Fisher Exact | 5.45E-02 | 1.99E-02 |
|  | CMC | 1.73E-01 | 4.68E-02 |
|  | SKAT | 1.25E-01 | 1.72E-02 |
|  | GSEA | 1.02E-01 | 6.06E-02 |
|  | PCA | 8.70E-01 | 3.20E-02 |

### Statistic genetics studies

Dr. Momiao Xiong, Professor, Division of Biostatistics, Human Genetics Center, at the School of Public Health at The University of Texas Health Science Center at Houston, is trained in genetic epidemiology and bioinformatics. As PI or co-Investigator on a number of NIH-funded grants, he has developed numerous statistical methods and computational algorithms for genetic studies of complex diseases involved in both qualitative and quantitative traits, DNA sequence analysis, detection of gene-gene interaction and gene-environment interaction, disease risk prediction and computational systems biology. Recently, his researches are to focus on developing methods for population and family-based association studies with next-generation sequencing data.

## Aim 1: To conduct exome sequencing of nuclear families to discover risk variants to SCZ in Chinese population

The primary goal of Aim 1 is to discover risk variants (both transmitted and de novo variants) for association testing in Aim 2. Towards this goal, we propose to conduct exome sequencing of 200 nuclear families with multiple affected individuals and one unaffected individual. In the literature, families with multiple affected individuals have been shown to enrich in variants, particulary rare variants, associated with the disorder. It is also shown that using multiple affected in the family could significant improve power in association analyses with both common and rare variants.32 With a family design, we can identify transmitted and de novo variants. Inclusion of an unaffected sibling can effectively exclude a large number of non-causal variants observed in in families, and significantly reduce the burden of association testing in Aim 2.

### Subject selection

We propose to use the Chinese family samples from the NIMH Genetics Initiative Repository (<https://www.nimhgenetics.org/>). We have applied and obtained approval to use the samples in NIH applications. There are two types of DNA samples available from Rutgers Coriell Repository (<http://ccr.coriell.org/>) for the Chinese samples. One is the DNA isolated from lymphoblastoid cell lines produced by the Epstein Barr Virus transfection. The other is cryopreserved lymphocytes. We plan to use DNA isolated from cell lines to conduct the first pass sequencing since cryopreserved lymphocytes are limited. We will use cryopreserved lymphocytes to isolate DNA to verify de novo mutation (see below).

Briefly, the samples were Chinese Han families collected by Dr. Ming Tsuang and his colleagues in Taiwan.74 The recruitment covers the entire Taiwan Island, and the inclusion criteria include a requirement of at least two DSM-IV affected individuals in a family, indicating all are multiplex families, i.e. all cases have positive family history of SCZ. A total of 607 families were recruited, including 1258 affected individuals. The mean age of onset was 22.6 ± 6.3. Of these families, 315 have both parents, and 292 have one parent. Based on the study32 of Dr. Momiao Xiong, one of our investigators, we plan to select all 179 families with one or both parents, two or more affected siblings and one unaffected sibling, with the unaffected sibling being examined as never mentally ill at age 32 or older (10 years older than the average age of onset in this collection). We will also include 21 families with both parents and multiple affected offspring. A total of 140 families and about 700 individuals will be selected. The rationales for these selections are: 1). The inclusion of both parents are necessary to discover de novo mutations in the affected offspring; 2). The inclusion of an unaffected sibling is to filter out a majority of non-causal variants found in the affected sibling. Since SCZ is a developmental disorder and most patients receive diagnosis in early 20s, we specify the age of unaffected individuals (never mentally ill) to be of 32 years or older to reduce the chance that these individual may develop SCZ in a late age; 3). For those families no eligible unaffected individuals, we will first select families with more affected individuals because these families have more power. These selection criteria aim to serve our goal of discovery of rare and de novo risk variants and to maximize our power in joint association analyses of both common and rare variants.

### Power estimate

140 family with two parents and three children (at least one child is ubaffected) and 5000 cases and 5000 controls. Five statistics: SFPCA, FPCA, CMC, VT and individual test (permutation was used to adjust for multiple test) were used to test association of a gene. We used the software ForSim software (

Lambert, B.W., Joseph, D., Terwilliger, J.D., and Weiss KM (2008). ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. Bioinformatics 24:1821–1822) to simulate pedigrees. Each individual in the pedigree has 189 variant sites with the MAF of all sites less than 0.01, but larger than 0.0001.Figure shows the power of tests as a function of proportion of risk variants under dominatn model . When we one quarter of risk variants we can reach more than 80% power.

Momiao, please so a power analysis using 140 families with two or more affected and one unaffected (total 700 subjects). You should include the replication sample of 5000 cases and 5000 controls.

### Exome sequencing

**Justification for BGI contract:** The technology for high throughput DNA sequencing has made significant progress in recent years, and it is now common practice to conduct exome sequencing at a reasonable cost. Exome sequencing starts with coding sequence extraction/enrichment, followed by library construction, and sequencing. In this application, we plan to subcontract the sequencing work to BGI, Beijing Genomics Institute, a worldwide leader in high throughput DNA sequencing market. One of the PIs of this application, Dr. Chen, has an on-going transcription sequencing project at BGI, and we have established a good working relationship with the people in BGI. For this application, we have obtained a favorable quote for exome sequencing at a cost of $650/subject. The subcontracted jobs include DNA sample preparation, library construction, DNA sequencing, genome alignment, gene mapping and SNV identification (see Appendix 1, BGI subcontract). The advantages of using the service of BGI are its experience in high throughput DNA sequencing, established pipeline of genomic analyses, sequence alignment, genome mapping and assembly, SNV calling and identification and cloud-computing support. Handling the large sequencing data and performing standard quality control, alignment to genomic reference, and identification of SNPs are challenging jobs, and require specialty training and experience. Since BGI has built the infrastructure and has expertise in these areas, we plan to include these tasks in the contract. It is our belief that contracting the sequencing and standard bioinformatics work to BGI will greatly increase our efficiency, reduce cost, and give us more time to focus on genetic analyses directly relevant to SCZ.

**Exome coverage:** Specifically, the DNA samples will be randomly fragmented into sizes of 150-300 base-pair length, coding sequences (including 5 KB promoter sequences) will be enrich and extracted with the SureSeletct II microarray from Agilent Company. These extracted DNA will be used to construct library for sequencing with Illimina’s HiSeq 2000 platform using paired-end chemistry. The sequencing will cover the coding sequences 50 times on average for each individual. Based on the estimate, this will enable us to detect 99.9% of variants with frequency of 0.5% or higher in the selected population.

**DNA Sequence Alignment and Polymorphism Identification:** The raw data produced from the sequencers will be quality-controlled using the pipeline procedures established at BGI. Specifically, we will use Illumina’s software to perform primary quality control for base calling and export sequence reads. The output sequence reads are then mapped to human genome reference. BGI has developed its own software (SOAP package)75 to handle downstream analyses of sequence reads, including alignment to genomic reference and identification of polymorphisms.76 This first pass variant identification data will used for further analyses to minimize the false calls (see below).

### Variant identification and characterization

Variant identification for NGS data is a challenge job. The 1000 Genomes Project and others have published methods for base-calling, sequence alignment and assembly, polymorphism identification for high throughput sequencing data.77-84 With these methods, the base-calling and SNP identification algorithms incorporate not only base-calling quality scores but also population data (such as human genome reference sequences and dbSNP data), linkage disequilibrium (LD) and reads from multiple individuals. These features minimize errors in SNP identification.85,86 Furthermore, transmission information among indiviudals in the pedigrees will also help to reduce base-calling errors. Based on our design, we will be able to discover transmitted and de novo variants, including SNVs, InDels and CNVs. Due to the limitation of the resources available to this RFA, we will focus on SNVs and large CNVs because the methods and protocols to identify these variants are relatively mature and stable. Other types of variants will be explored when resources allow.

**SNV identification:** We plan to use the GATK package87 to refine SNV identification. Specifically, we will export the SAM files for each individual from the SOAP software,61,75 and feed these files for variant discovery using reads from all sequenced individuals. We will adopt the algorithms developed by DePristo et al86 for local multiple sequence alignment, base-calling quality recalibration, multi-sample SNP calling and SNP quality score recalibration. In this process, pedigree information, known polymorphic sites for Chinese from the dbSNP and 1000 Genomes Project will be incorporated into the analyses. A final evaluation of overall SNV calling quality for the entire dataset will be assessed by the transition and transversion ratio. Based on a previous study, a transition/transversion ratio of 3.0-3.5 is expected for coding sequences.88 A significant deviation from this expectation is an indication of high rate of false SNV calling, and further data-cleaning would be needed and SNV calling redone.

**CNV identification:** CNV discovery is more challenge with exome sequencing data than that with whole genome sequencing data, mainly due to uneven coverage of exons. For whole genome sequencing, CNV discovery is largely determined by the depth of coverage, assumptions used in whole genome sequencing may not hold in exome sequencing.89 Therefore, we will use the CONTRA software90 that is designed specifically for CNV identification in targeted sequencing and exome sequencing. This software uses multiple samples to create a baseline for reference, taking into account of the library size, total number of reads, reading length, and percentage on target for each individual sample. A base-level log ratio is calculated with the adjusted coverage for each sample, and the mean of base-level log ratio over a specified region will be used to score CNVs. We like the feature of this package that allows user to specify selected samples and region as reference for CNV scoring. Realizing that CNV calling can be problematic, we will use another software package, ExomeCNV,89 to re-score the samples. Only concordant CNVs from the two packages are selected for association testing.

**SNV selection and filtering:** In this study, we sequence 4-5 individuals in a family. As demonstrated in a recent study,91 using family information to phase individual’s haplotypes can identify > 90% sequencing errors and further improve SNV identification. We plan to adopt the same strategy, and use family information to phase out the haplotypes of individuals within a family, and correct sequencing errors accordingly. A hypothesis that cases with FH is that these families carry private mutations. With sequencing data from 2 or more affected individuals and 1 unaffected individual in a family, we will be able to identify a group of private mutations observed only in the affected individuals in the families. For these private mutations, we can impose further selections. We can select concordant variants observed only in affected individuals in a family, we can also require these same variants observed in at least one more other families, or the variants occurred in the same genes in different families. Here we will take advantage of the existing large exome- and whole-genome sequencing data available from dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) and 1000 Genomes Project (<http://www.1000genomes.org/>) to further filter out those variants observed in mentally healthy individuals. Since the databases contain substantially large number of healthy individuals that can be considered as controls for schizophrenia studies, this will significantly increase our statistic power to detect risk variants. We can characterize the remaining mutations into different categories such as synonymous, non-synonymous, stop codon, splicing variant, and frame-shifting variant. For non-synonymous variants, potential consequences can be evaluated with the MAPP,92,93 PolyPhen,94-96 SIFT97 and PANTHER98-100 programs. For these mutations, we can devise different tests to evaluate their association with SCZ. We can conduct a binomial test to evaluate whether any of these types of variants are enriched in cases compared to healthy controls, or we can integrate these mutations with other variants to jointly test their effects across a gene or a biological pathway (see below). Individual variants/gene-based testing

**De novo mutation verification:**  Our design allows us to identify de nova mutations by comparing the variants observed between parents and affected offspring. Because the DNA samples we plan to use are from lymphoblastoid cell lines, some of the variants observed only in offspring may be false due to the process of cell line development (Epstein Barr Virus transfection). To exclude these false positives, we will obtain cryopreserved lymphocytes from Coriell Repository, extract DNA, and resequence these samples to verify those de novo mutations. The verified de novo mutations will be tested as other SNVs described above.

### Association analyses

In addition to the enrichment test for functional and de novo mutations, we will conduct association analyses using all variants discovered in the sequencing. In these analyses all variants found in a gene or a biological pathway are combined and tested using the new statistics developed by Dr. Momiao Xiong, our investigator from UT Health Center at Houston.

**A unified general framework for sequence-based association studies with pedigree structure and unrelated individuals.** We outline the general procedures to extend population-based association tests for NGS to a general case with multiple families and unrelated individuals. Consider the problem of testing for the association of a genomic region in a sample of n sequenced individuals from multiple families or unrelated individuals with unknown population structures. We assume that the number of variants in a genomic region, or number of functional principal components or the number of collapsed summary variables is k. For each individual, we define a vector of variables , . Define a case indicator vector , whose i-th entry is 1 if individual *i* is a case and 0 if individual is a control, and , a column vector of 1s of length . Define , where is the number of affected individuals,  is a dimensional identity matrix, denotes the Kronecker product of two matrices. Similar to population-based association studies where we compare the difference in frequencies of alleles or collapsing variables between cases and controls, we compare the same differences in a genomic region but between affected and unaffected individuals in pedigrees or populations. The statistic is defined as

, where 

A key step is to calculate the covariance matrix of the variables  Our innovative approach is to decompose the covariance matrix  into a Kronecker product of the covariance matrix of the variables Z for outbred individuals and a kinship matrix accounting for relatedness, inbreeding, and population structure8: , which implies\* A remarkable feature is that the test statistic can be simplified to a product of the corresponding population-based statistics and a correction factor: where  and are the average of the vector Z for affected and unaffected individuals,  is the correction factor to be applied to the statistic to have a valid test in the presence of pedigree structures and cryptic relatedness. Under the null hypothesis of no association,  is distributed as a central distribution.

**Family-based generalized, Collapsing methods, CMC, WSS, VT, SKAT, SFPCA statistic, and other sequence-based association test statistics.**

It is increasingly recognized that analyzing samples from populations and pedigrees separately is highly inefficient. It is natural to unify population and family study designs for association studies. However, the major critical barriers are lack of the statistical methods for sequence-based association studies with extended pedigrees and for integrating the association results across the study designs. Since the statistical methods appropriate for either design differ substantially, it is difficult to integrate the association results across the sample designs6. To overcome these problems, we extend the population based association tests to any pedigrees or mixed pedigrees and unrelated individuals. Specifically, the generalized test103, multivariate distance matrix regression104 and kernel based association test 105, the Weighted Sum-square (WSS) [5], the variable-threshold (VT) test [7]; the Sequence Kernel Association Test (SKAT) [8] which are originally designed for population-based association studies with NGS data can be extended to family-based association studies as described in the previous section (Shugart et al. 2012). Recently, Dr. Xiong has also developed the FPCA and SFPCA methods to jointly test common and rare variants discovered by NGS in a gene or a genomic region. The FPCA (Luo et al. 2011) and SFPCA methods can also be generalized to samples with arbitrary combinations of related (pedigrees) and unrelated individuals by the principal described in the previous section and have been published (Zhu and Xiong 2012). The FPCA and SFPCA methods can be used to test the association of the entire allelic spectrum of genetic variation. The joint genetic analysis of related and unrelated individuals can effectively use the observed transmission information in the pedigrees and the linkage disequilibrium (LD) information hidden in the history of populations, and effectively use rare and common variants segregating in the pedigrees and the population. Our approach estimates both within and between family information from sequence data to simultaneously correct for family structure and cryptic relatedness while maintaining high power.

**Analysis of rare variants:** Due to the low frequency of rare variants, testing these variants individually, such as the χ2 test, Fisher’s exact test, Cochran-Armitage trend test and logistic regression, may be underpowered.101,102Therefore, it is a necessity to jointly test multiple variants at a gene, given locus/genomic interval. There are two basic strategies for locus based testing. One is a multivariate method that combines information across multiple variants at a given locus and test them simultaneously. A multivariate generalized test,103 proposed by our investigator Dr. Momiao Xiong, is one of the methods. Other methods include multivariate distance matrix regression104 and kernel based association test.105 The second strategy is to collapse rare variants in an individual across a locus by frequency and potentially functional consequences as measured by MAPP,92,93 PolyPhen,94-96 SIFT97 and PANTHER98-100 programs. Additionally, the quality score of the variants can be incorporated in the analyses as well.106 These methods for testing rare variants, in general, are less powerful than the methods jointly test both common and rare variants together (see preliminary data and description below).

**Application of the functional principal component analysis to jointly test common and rare variants in family samples:** Recently, Dr. Xiong has developed the FPCA and SFPCA methods to jointly test common and rare variants discovered by NGS in a gene.32,63 These methods have better power than methods that test individual variants or simply clamp rare variants together. In this application, we extend the statistics to a general framework where multiple families and additional population structures are presented in the samples. Specifically, let  be a genotypic function of the *i-*the individual, where *t* is the genomic position. The smoothed functional component scores is defined as , is a penalty parameter and is orthornomal eigen functions. Our purpose is to use the functional principal component scores to develop test statistics that can be applied to pedigrees. To achieve this, we first calculate the covariance matrix of the functional principal component scores. Let where is the number of principal components. We assume that the sampling covariance matrix within individual is estimated by . Let be the vector of averages of the functional principal component scores in cases and controls, respectively, and be the number of sampled individuals in cases. The test statistic is then given by Under the null hypothesis of no association of the genomic region, the statistic will be asymptotically distributed as a central distribution where is the number of functional principal components in the eigen equation expansion of genotypic functions.

**The smoothed functional principal component analysis for pathway association:** The FPCA for testing the association of a gene can be extended to testing association of a pathway by extending FPCA from an interval to multiple intervals. Consider a linear combination of functional values:, where is a weight function and  is a centered indicator function for genotypes. To capture the genetic variations in the genotype functions, we chose weight function  to maximize the variance of  To improve the smoothness of the estimated functional principal component curves, we impose the roughness penalty on functional principal component weight functions. The smoothed functional principal components (SFPCA)can be obtained by solving the integral equations:where are the functional principal components. Let be the functional principal component scores of and . Let  and , S be the pooled covariance matrix of the functional principal component scores. Define Then, the statistic is defined as . Under the null hypothesis of no association of a pathway , the statistic  is a central distribution. The population-based FPCA and SFPCA for pathway association test can also be extended to pedigrees.

### Data sharing plan

The sequencing data for the Chinese families will be shared in compliance with NIH policies. Since the pedigrees we used will be from NIMH repository, personal identification information has already been removed, we will organize the sequencing and appropriate phenotypic data and transport to dbGaP for sharing. Since the exome sequencing data can be used for personal identification, to protect the research subjects, we will put in explicit prohibition for using the data for personal identification, and all requests for data access will need IRB approval. The details of when the data will be shared and what to be shared will be negotiated with NIMH officials following NIH guidelines.

## Aim 2: Verification of top candidates identified from DNA sequencing

The purpose of this aim is to verify the promising candidates discovered in Aim 1 above. As stated above, exome sequencing data can be used to discover CNVs, but due to the limitation of this RFA, we will study only SNVs in this aim. The sequencing will produce a substantially large number of SNVs. A key issue for replication will be variant selection. Based on our design, we will be able to detect transmitted variants (both common and rare) and de novo mutations, which can be classified as rare variants as well. In the last few years, we have learned that the likelihood of finding causative variants with modest frequencies (say, 1% or greater) is not good,107 we will focus on rare variants with frequencies < 1%. With 5,000 cases and 5,000 controls, the power to detect individual variants with such frequencies is limited. Therefore, we need to use innovative statistics to jointly test multiple rare (and common) variants.

### Variant selection for replication

The criteria for selecting rare variants for replication in a large case control sample (see below) include:

1. Ranked on the top 10% of candidates based on the testing described in Aim 1;
2. Observed in multiple families;
3. Observed in cases only;
4. Preferably functional variants (stop codon, missense mutation, frame-shifting/splicing variants etc.);
5. Preferably in the same gene or known pathway with multiple other top-ranking candidates from Aim 1 or from literature/public databases.

### Replication samples and genotyping

Our Chinese team has published a large GWAS with a combined sample size of 8,123 cases and 11,007 controls (3,750 cases and 6,468 controls in discovery sample, and 4,383 cases and 4,539 controls in replication sample).36 We plan to use these subjects to verify findings from the sequencing of families in Aim 1. (Dr. He, please indicate whether all of these samples can be used. I used conservative numbers of 5000 cases and 5000 controls. If we can use all of them, we should say so to increase our power. It will be good that if you can write a letter to certify that these samples will be available for this study) Since the Chinese government has policies on DNA exporting, we will conduct the replication study in China at the site of our Chinese collaborator, i.e. the Bio-X Center of Shanghai Jiaotong University. Up to 100 top selected candidates will be genotyped for these samples using the TaqMan method.108 Because the variants to be typed have very low frequencies, genotyping error in a few individuals would have severe consequences. Therefore, all heterozygotes and minor allele homozytes will be visually inspected and retyped to confirm.

### Data analysis

We plan to conduct combined analyses for the sequencing and replication samples. Since the number of variants tested in the replication samples is limited, multiple testing correction should be set as in the Aim 1 for the total number of variants tested. As outline above, we will conduct binomial/Fisher’s exact test for individual variants after combining the samples, and we also conduct gene- and pathway-based analyses, jointly testing all variants (both common and rare) together.

## Aim 3: Comparative analyses of genetic structure between the Chinese and Caucasian population

Our current understanding of the genetics of SCZ is largely derived from the studies of Caucasian populations. From the literature, there are hints that some risks to SCZ are shared across populations while others may be ethnic specific. We have access to the largest SCZ GWAS dataset for the Caucasians, the PGC SCZ dataset,4 and our Chinese collaborator has recently published a large GWAS of Chinese samples.36 Furthermore, there are several ongoing large SCZ sequencing studies for Caucasians, and the study proposed in this application will be one of the largest sequencing project in Chinese population. With these data, there are opportunities to systematically examine the shared and ethnic specific risk factors between the two populations.

### Testing the shared genetic factors between the Chinese and Caucasian populations

**GWA data imputation:** To facilitate the study of genetic risk correlation, we plan to first conduct imputation to the same reference panel, so both the Chinese and Caucasian GWAS datasets have the same makers covered across the genome. Since we intend to study the shared genetic risks, we need to use a reference panel containing individuals from both populations, because there are regions in the genome that Chinese and Caucasian have different LD patterns. For this purpose, we plan to use the data from the 1000 Genomes Project, which includes 200 Chinese and 200 Caucasians. The BEAGLE program109-112 will be used for this imputation.

**Evaluation of the correlation of genetic risks between Chinese and Caucasian population:** Once we have the GWA data imputed for both Chinese and Caucasian samples, we can ask the question whether and to what extent the genetic risk between the two populations correlates. We propose to use the CGAT program113 to calculate genetic risk scores for the Caucasian and use this genetic risk score to predict the affection status of the Chinese samples by standard logistic regression. We can also reverse the process and use genetic risk score in Chinese to predict the affection status in the Caucasians. These analyses will provide evidence whether and how much the genetic risks are shared between the two populations. Based on recent studies of other complex diseases,114 we would expect that there are significant overlaps in genetic risk factors, and we would also expect that the genetic risk scores calculated from the Caucasians are a better predictor because the PGC SCZ GWAS sample size is much larger than the Chinese GWAS, and therefore it captures more risk loci than the Chinese GWAS does.

**Identification of shared loci, genes and biological pathways:** The assumption that common genetic risk factors exist across ethnicities can be made at different levels. The first one is that all human races have common origin and share common evolutionary history. If the causative mutations predate the divergence of the ethnic groups under investigation, the shared evolutionary history, or LD patterns, between the populations in question, constitutes the base for shared liability. Under this condition, the shared risk alleles should have similar frequencies as these are a reflection of LD in the regions. Secondly, the same disease has same/similar pathophysiology, and involves genes in certain functions and biological pathways. Conceptually, independent mutations can occur in these same genes and pathways in different populations after diverging from the common ancestry. Under this condition, the risk alleles can be same (recurrent mutations) or different, and the LD pattern in the region can be different as well. With these rationales, for risk alleles originated from common ancestry, we need statistic power to overcome sampling bias (winner’s curse) to verify findings in one population by another population. Since the PGC SCZ GWA has larger sample size, it is likely that we have more true signals in this dataset. To maintain the power, we propose to test only those markers with p values ≤ 5x10-5 in the PGC datast. Furthermore, we will restrict the test to markers with LD of r2 ≥ 0.8 between the Chinese and Caucasian populations from the data of 1000 Genomes Project. With these selections, the Chinese GWA will have sufficient power to test 1000 markers, and any markers with p ≤ 5x10-5 would be deemed statistically significant, or the variants are potentially shared liability between the two populations.

For those shared liabilities converged on gene functions and biological pathways through independent mutations, the test needs to be done in the unit of gene or pathway. With the gene-based approach, all markers in the intragenic and regulatory regions of a gene should be considered. Since the boundary of a gene is somewhat arbitrary, we plan to adopt a definition based on the study of expression quantitative trait locus,115 including 100 kb upstream and 40 kb downstream of the coding sequence a gene. There are several strategies116-118 for gene-based association analyses, we plan to use the methods our collaborator Dr. Xiong developed,32,63 because these ware designed to integrate both common and rare variants for association testing and were more powerful than those methods simply clamping variants together.

### Testing ethnic specific risk factors

Most studies using a single population inexplicitly target ethnic specific risk variants. Whether the risk variants found in such studies are truly ethnic specific, verification in other populations are necessary. Based on the sample size of the PGC SCZ GWA study (9,394 cases and 12,462 controls) and Chinese GWA study (3,750 cases and 6,468 controls), we can only test a few top candidates from each of population to evaluate if these loci are ethnic specific. The main concern here is the replication power. Only if a locus found in one population cannot be replicated with a sample of sufficient power in another population, we can be reasonably confident to conclude that the locus is ethnic specific. With this rationale, we will select top candidates from each GWA study, and select only those candidates that can be replicated given the observed OR in one population and the replication sample size in the other population. In addition, we will impose a restriction that the selected loci need in a region under differential selection during evolution between Chinese and Caucasian. Specifically, we will select the top candidates from the Chinese GWA study, conduct power calculation using the observed ORs in Chinese and the PGC sample, and vice versa. We will then calculate the fixation index, FST, a measure of population differentiation, for these candidates between the two populations using the locus-specific approaches.119-121 This locus specific FST is then compared to the population mean FST across the genome between Caucasian and Chinese. If the locus specific FST falls outside the 95% confidence intervals of the population mean, the locus is defined as differentially selected between the populations, and is, therefore, included from further testing.

In this application, we seek to understand the genetic architecture in Chinese population by sequencing 140 multiplex pedigrees followed by verification with 5,000 cases and 5,000 controls. Recent studies with Caucasian populations have demonstrated that rare variants, including CNVs, may be an important source of genetic risk in SCZ. What roles rare variants play in Chinese population are not clear. We take a family design that includes parents and affected and unaffected siblings. With this design, we will be able to discover both rare and de nova risk variants transmitted to affected individuals. Testing these variants in a much larger case control sample will allow us to verify if these rare variants are associated with SCZ and have a role in the genetic architecture in Chinese population. Our design is novel, and we have sufficient power to detect and verify at least some variants. In the preliminary results, we have documented our experience and expertise (i.e. genetic studies of SCZ, NGS sequencing and association analyses of rare and common variants) directly relevant to this application. We are confident that we will be able to discover new risk variants in the Chinese samples and provide new insights on our understanding of the genetics of SCZ.

References

1. O'Donovan MC, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I et al. Analysis of 10 independent samples provides evidence for association between schizophrenia and a SNP flanking fibroblast growth factor receptor 2. *Mol Psychiatry* 2009; **14:** 30-36.

2. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 2008; **40:** 1053-1055.

3. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460:** 748-752.

4. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA et al. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011; **43:** 969-976.

5. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D et al. Common variants conferring risk of schizophrenia. *Nature* 2009; **460:** 744-747.

6. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460:** 753-757.

7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461:** 747-753.

8. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* 2011; **35:** 310-317.

9. Crow TJ. 'The missing genes: what happened to the heritability of psychiatric disorders?'. *Mol Psychiatry* 2011; **16:** 362-364.

10. Gershon ES, Alliey-Rodriguez N, Liu C. After GWAS: searching for genetic risk for schizophrenia and bipolar disorder. *Am J Psychiatry* 2011; **168:** 253-256.

11. Grozeva D, Kirov G, Ivanov D, Jones IR, Jones L, Green EK et al. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* 2010; **67:** 318-327.

12. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008; **320:** 539-543.

13. Ingason A, Rujescu D, Cichon S, Sigurdsson E, Sigmundsson T, Pietilainen OP et al. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry* 2011; **16:** 17-25.

14. Karlsson R, Graae L, Lekman M, Wang D, Favis R, Axelsson T et al. MAGI1 copy number variation in bipolar affective disorder and schizophrenia. *Biol Psychiatry* 2012; **71:** 922-930.

15. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 2012; **17:** 142-153.

16. Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP et al. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 2010; **87:** 618-630.

17. Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC et al. Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 2010; **87:** 229-236.

18. Rees E, Kirov G, O'Donovan MC, Owen MJ. De novo mutation in schizophrenia. *Schizophr Bull* 2012; **38:** 377-381.

19. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR et al. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* 2010; **87:** 316-324.

20. Gauthier J, Champagne N, Lafreniere RG, Xiong L, Spiegelman D, Brustein E et al. De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. *Proc Natl Acad Sci U S A* 2010; **107:** 7863-7868.

21. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 2011;

22. Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 2011; **72:** 951-963.

23. Tarabeux J, Champagne N, Brustein E, Hamdan FF, Gauthier J, Lapointe M et al. De novo truncating mutation in Kinesin 17 associated with schizophrenia. *Biol Psychiatry* 2010; **68:** 649-656.

24. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 2011; **43:** 864-868.

25. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* 2011; **13:** 135-145.

26. Owen MJ, Craddock N, O'Donovan MC. Suggestion of roles for both common and rare risk variants in genome-wide studies of schizophrenia. *Arch Gen Psychiatry* 2010; **67:** 667-673.

27. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry* 2011;

28. Shi G, Rao DC. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genet Epidemiol* 2011; **35:** 572-579.

29. Ionita-Laza I, Ottman R. Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 2011; **189:** 1061-1068.

30. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. *Am J Hum Genet* 2012; **90:** 7-24.

31. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res* 2011; **21:** 1099-1108.

32. Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. *Am J Hum Genet* 2012; **90:** 1028-1045.

33. Agerbo E, Mortensen PB, Wiuf C, Pedersen MS, McGrath J, Hollegaard MV et al. Modelling the contribution of family history and variation in single nucleotide polymorphisms to risk of schizophrenia: a Danish national birth cohort-based study. *Schizophr Res* 2012; **134:** 246-252.

34. Francks C, Tozzi F, Farmer A, Vincent JB, Rujescu D, St CD et al. Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. *Mol Psychiatry* 2010; **15:** 319-325.

35. Tandon R, Keshavan MS, Nasrallah HA. Schizophrenia, "Just the Facts" What we know in 2008. 2. Epidemiology and etiology. *Schizophr Res* 2008; **102:** 1-18.

36. Shi Y, Li Z, Xu Q, Wang T, Li T, Shen J et al. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat Genet* 2011;

37. Chen Q, Che R, Wang X, O'Neill FA, Walsh D, Tang W et al. Association and expression study of synapsin III and schizophrenia. *Neurosci Lett* 2009; **465:** 248-251.

38. Chen X, Sun C, Chen Q, O'Neill FA, Walsh D, Fanous AH et al. Apoptotic engulfment pathway and schizophrenia. *PLoS ONE* 2009; **4:** e6875-

39. Kendler KS, Gruenberg AM. An independent analysis of the Danish Adoption Study of Schizophrenia. VI. The relationship between psychiatric disorders as defined by DSM-III in the relatives and adoptees. *Arch Gen Psychiatry* 1984; **41:** 555-564.

40. Kendler KS, Adler D. The pattern of illness in pairs of psychotic siblings. *Am J Psychiatry* 1984; **141:** 509-513.

41. Kendler KS, Masterson CC, Ungaro R, Davis KL. A family history study of schizophrenia-related personality disorders. *Am J Psychiatry* 1984; **141:** 424-427.

42. Kendler KS, Masterson CC, Davis KL. Psychiatric illness in first-degree relatives of patients with paranoid psychosis, schizophrenia and medical illness. *Br J Psychiatry* 1985; **147:** 524-531.

43. Kendler KS, McGuire M, Gruenberg AM, O'Hare A, Spellman M, Walsh D. The Roscommon Family Study. III. Schizophrenia-related personality disorders in relatives. *Arch Gen Psychiatry* 1993; **50:** 781-788.

44. Williams HJ, Norton N, Dwyer S, Moskvina V, Nikolov I, Carroll L et al. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Mol Psychiatry* 2010;

45. Shi Y, Zhao X, Yu L, Tao R, Tang J, La Y et al. Genetic structure adds power to detect schizophrenia susceptibility at SLIT3 in the Chinese Han population. *Genome Res* 2004; **14:** 1345-1349.

46. Zhao X, Tang R, Gao B, Shi Y, Zhou J, Guo S et al. Functional variants in the promoter region of Chitinase 3-like 1 (CHI3L1) and susceptibility to schizophrenia. *Am J Hum Genet* 2007; **80:** 12-18.

47. Li J, Zhou G, Ji W, Feng G, Zhao Q, Liu J et al. Common variants in the BCL9 gene conferring risk of schizophrenia. *Arch Gen Psychiatry* 2011; **68:** 232-240.

48. Chen Q, He G, Qin W, Chen QY, Zhao XZ, Duan SW et al. Family-based association study of synapsin II and schizophrenia. *Am J Hum Genet* 2004; **75:** 873-877.

49. Fan JB, Tang JX, Gu NF, Feng GY, Zou FG, Xing YL et al. A family-based and case-control association study of the NOTCH4 gene and schizophrenia. *Mol Psychiatry* 2002; **7:** 100-103.

50. Li D, He L. Association study between the dystrobrevin binding protein 1 gene (DTNBP1) and schizophrenia: a meta-analysis. *Schizophr Res* 2007; **96:** 112-118.

51. Li D, Collier DA, He L. Meta-analysis shows strong positive association of the neuregulin 1 (NRG1) gene with schizophrenia. *Hum Mol Genet* 2006; **15:** 1995-2002.

52. Li J, Liu J, Feng G, Li T, Zhao Q, Li Y et al. The MDGA1 gene confers risk to schizophrenia and bipolar disorder. *Schizophr Res* 2011; **125:** 194-200.

53. Lo WS, Lau CF, Xuan Z, Chan CF, Feng GY, He L et al. Association of SNPs and haplotypes in GABAA receptor beta2 gene with schizophrenia. *Mol Psychiatry* 2004; **9:** 603-608.

54. Tang JX, Zhou J, Fan JB, Li XW, Shi YY, Gu NF et al. Family-based association study of DTNBP1 in 6p22.3 and schizophrenia. *Mol Psychiatry* 2003; **8:** 717-718.

55. Chen X, Wang X, Hossain S, O'Neill FA, Walsh D, Pless L et al. Haplotypes spanning SPEC2, PDZ-GEF2 and ACSL6 genes are associated with schizophrenia. *Hum Mol Genet* 2006; **15:** 3329-3342.

56. Chen X, Wang X, Chen Q, Williamson V, van den Oord E, Maher BS et al. MEGF10 association with schizophrenia. *Biol Psychiatry* 2008; **63:** 441-448.

57. Chen X, Wang X, Hossain S, O'Neill FA, Walsh D, van den Oord E et al. Interleukin 3 and schizophrenia: the impact of sex and family history. *Mol Psychiatry* 2007; **12:** 273-282.

58. Chen Q, Wang X, O'Neill FA, Walsh D, Fanous A, Kendler KS et al. Association study of CSF2RB with schizophrenia in Irish family and case - control samples. *Mol Psychiatry* 2008; **13:** 930-938.

59. Chen X, Lee G, Maher BS, Fanous AH, Chen J, Zhao Z et al. GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia. *Mol Psychiatry* 2011; **16:** 1117-1129.

60. Xu J, Sun J, hen J, ang L, elm M, Dubovsky SL et al. RNA-seq analysis implicates dysregulation of the immune system in schizophrenia. *BMC Genomics* 2012;

61. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009; **25:** 1966-1967.

62. Niu B, Zhu Z, Fu L, Wu S, Li W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 2011; **27:** 1704-1705.

63. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res* 2011;

64. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5:** e1000384-

65. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86:** 832-838.

66. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83:** 311-321.

67. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. *Eur J Hum Genet* 2010; **18:** 1045-1053.

68. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89:** 82-93.

69. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102:** 15545-15550.

70. Giacca M, Zacchigna S. VEGF gene therapy: therapeutic angiogenesis in the clinic and beyond. *Gene Ther* 2012; **19:** 622-629.

71. Bing RJ. Myocardial ischemia and infarction: growth of ideas. *Cardiovasc Res* 2001; **51:** 13-20.

72. Dobaczewski M, Chen W, Frangogiannis NG. Transforming growth factor (TGF)-beta signaling in cardiac remodeling. *J Mol Cell Cardiol* 2011; **51:** 600-606.

73. Bujak M, Frangogiannis NG. The role of TGF-beta signaling in myocardial infarction and cardiac remodeling. *Cardiovasc Res* 2007; **74:** 184-195.

74. Hwu HG, Faraone SV, Liu CM, Chen WJ, Liu SK, Shieh MH et al. Taiwan schizophrenia linkage study: the field study. *Am J Med Genet B Neuropsychiatr Genet* 2005; **134B:** 30-36.

75. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008; **24:** 713-714.

76. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 2009; **19:** 1124-1132.

77. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467:** 1061-1073.

78. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008; **18:** 1851-1858.

79. Kao WC, Song YS. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *J Comput Biol* 2011; **18:** 365-377.

80. Kao WC, Stevens K, Song YS. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res* 2009; **19:** 1884-1895.

81. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008; **18:** 810-820.

82. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009; **19:** 1117-1123.

83. Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, Morris RW. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* 2010; **26:** 2803-2810.

84. Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ et al. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res* 2010; **20:** 537-545.

85. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011; **12:** 443-451.

86. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43:** 491-498.

87. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20:** 1297-1303.

88. Ebersberger I, Metzler D, Schwarz C, Paabo S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* 2002; **70:** 1490-1497.

89. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011; **27:** 2648-2654.

90. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012; **28:** 1307-1313.

91. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* 2011; **7:** e1002280-

92. Binkley J, Karra K, Kirby A, Hosobuchi M, Stone EA, Sidow A. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. *Genome Res* 2010; **20:** 142-154.

93. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 2005; **15:** 978-986.

94. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002; **30:** 3894-3900.

95. Sunyaev S, Ramensky V, Koch I, Lathe W, III, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* 2001; **10:** 591-597.

96. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000; **16:** 198-200.

97. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31:** 3812-3814.

98. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005; **33:** D284-D288.

99. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 2003; **13:** 2129-2141.

100. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* 2003; **31:** 334-341.

101. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 2011; **35 Suppl 1:** S12-S17.

102. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010; **44:** 293-308.

103. Xiong M, Zhao J, Boerwinkle E. Generalized T2 test for genome association studies. *Am J Hum Genet* 2002; **70:** 1257-1268.

104. Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* 2006; **79:** 792-806.

105. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 2010; **34:** 213-221.

106. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum Hered* 2012; **73:** 84-94.

107. Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D, Maia JM et al. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet* 2012; **91:** 303-312.

108. Livak KJ. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal* 1999; **14:** 143-149.

109. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84:** 210-223.

110. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* 2009; **85:** 847-861.

111. Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet* 2010; **86:** 526-539.

112. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* 2008; **124:** 439-450.

113. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 2011; **43:** 519-525.

114. Fu J, Festen EA, Wijmenga C. Multi-ethnic studies in complex traits. *Hum Mol Genet* 2011; **20:** R206-R213.

115. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008; **4:** e1000214-

116. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 2010; **87:** 139-145.

117. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet* 2011; **88:** 283-293.

118. Li M, Wang K, Grant SF, Hakonarson H, Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* 2009; **25:** 497-503.

119. Garte S. Locus-specific genetic diversity between human populations: an analysis of the literature. *Am J Hum Biol* 2003; **15:** 814-823.

120. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 2010; **10:** 564-567.

121. Browning SR, Weir BS. Population structure with localized haplotype clusters. *Genetics* 2010; **185:** 1337-1344.

[5] Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistics.** *Plos Genet* 2009, **5:**e1000384.

[7] Price AL, Kryukov GV, Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86:** 982.

[8] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare variant association testing for sequencing data using the sequence kernel association test (SKAT).** *Am J Hum Genet* 2011, **89:** 82-93.

Luo L, Boerwinkle E, Xiong M: **Association studies for next-generation sequencing. *Genome Res*** 2011, **21:** 1099-1108.

Shugart Y, Zhu Y, GuoW, Xiong MM. (2012) Weighted Pedigree-based Statistics for Testing the Association of Rare Variants. BMC Genomics (Revised).

Zhu Y and Xiong MM (2012) Family-Based Association Studies for Next-Generation Sequencing. Am J Human Genet. 90(6):1028-1045.